SOUNDCOMPASS: NAVIGATING TARGET SOUND EXTRACTION WITH EFFECTIVE DIRECTIONAL CLUE INTEGRATION IN COMPLEX ACOUSTIC SCENES

Dayun Choi[1] and Jung-Woo Choi[1]*

[1]School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
{cdy3773, jwoo}@kaist.ac.kr

ABSTRACT

Recent advances in target sound extraction (TSE) utilize directional clues derived from direction of arrival (DoA), which represent an inherent spatial property of sound available in any acoustic scene. However, previous DoA-based methods rely on hand-crafted features or discrete encodings, which lose fine-grained spatial information and limit adaptability. We propose SoundCompass, an effective directional clue integration framework centered on a Spectral Pairwise INteraction (SPIN) module that captures cross-channel spatial correlations in the complex spectrogram domain to preserve full spatial information in multichannel signals. The input feature expressed in terms of spatial correlations is fused with a DoA clue represented as spherical harmonics (SH) encoding. The fusion is carried out across overlapping frequency subbands, inheriting the benefits reported in the previous band-split architectures. We also incorporate the iterative refinement strategy, chain-of-inference (CoI), in the TSE framework, which recursively fuses DoA with sound event activation estimated from the previous inference stage. Experiments demonstrate that SoundCompass, combining SPIN, SH embedding, and CoI, robustly extracts target sources across diverse signal classes and spatial configurations.

***Index Terms*—** directional clue, target sound extraction, spectral pairwise interaction, spherical harmonics, iterative refinement.

1. INTRODUCTION

Target sound extraction (TSE) [1] refers to the task of selectively extracting a target audio source from a complex acoustic scene. TSE has gained increasing attention due to its wide range of practical applications in hearing aids [2], augmented/virtual reality (AR/VR) [3], and teleconferencing [4]. In these scenarios, isolating a desired source from interfering signals and background noise is critical for both human perception and machine-based recognition.

Recent studies in deep learning have investigated TSE using auxiliary clues that guide the model toward the target source. Illustrative auxiliary clues encompass class labels [5, 6], text descriptions [7, 8], visual cues [9], or their combinations [10]. In addition to these, the direction of arrival (DoA) has been utilized as a notable clue that leverages the spatial characteristic to isolate a target from interfering sources, irrespective of temporal or spectral attributes.

The effective use of DoA clues hinges not only on selecting input features that describe the spatial aspects of multichannel signals but also on the clue-fusion architecture in which DoA clues are articulated and integrated with these input features. With respect to input features, prior work [11, 12] has focused on manually designed features such as inter-channel phase differences (IPD) or inter-channel level differences (ILD), which improve TSE performance compared to using raw waveforms or complex spectrogram input. Nevertheless, they might lose essential spatial information, and whether such features are the optimal choice for capturing spatial relationships remains an unresolved issue.

Beyond the choice of input features, prior studies have also differed in how DoA clues are represented and fused with these features. Some studies [11–13] used IPD and target phase difference (TPD) computed from target DoA and known microphone positions. Other approaches [14–16] adopted one-hot or binary encodings, which ignore the continuous and periodic nature of angular space by treating adjacent directions as independent categories, leading to increased input dimensionality and hindering generalization to unseen or intermediate directions. In these approaches, DoA clues were combined with input features through operations such as multiplication [14], initial recurrent states [15], or attention keys/values [16]. More recently, M2M-TSE [17] and DSENet [18] employed cyclic positional (cyc-pos) embeddings that explicitly capture the periodic structure of angular space. M2M-TSE further applied them by broadcasting across time only when the target is active and multiplying them with input features. Since the exact duration of the target activity is generally unknown, this approach needs to handle temporal uncertainty when incorporating directional clues.

To address these limitations, we propose SoundCompass, a framework for effective directional clue integration. The key contributions of the proposed framework are as follows:
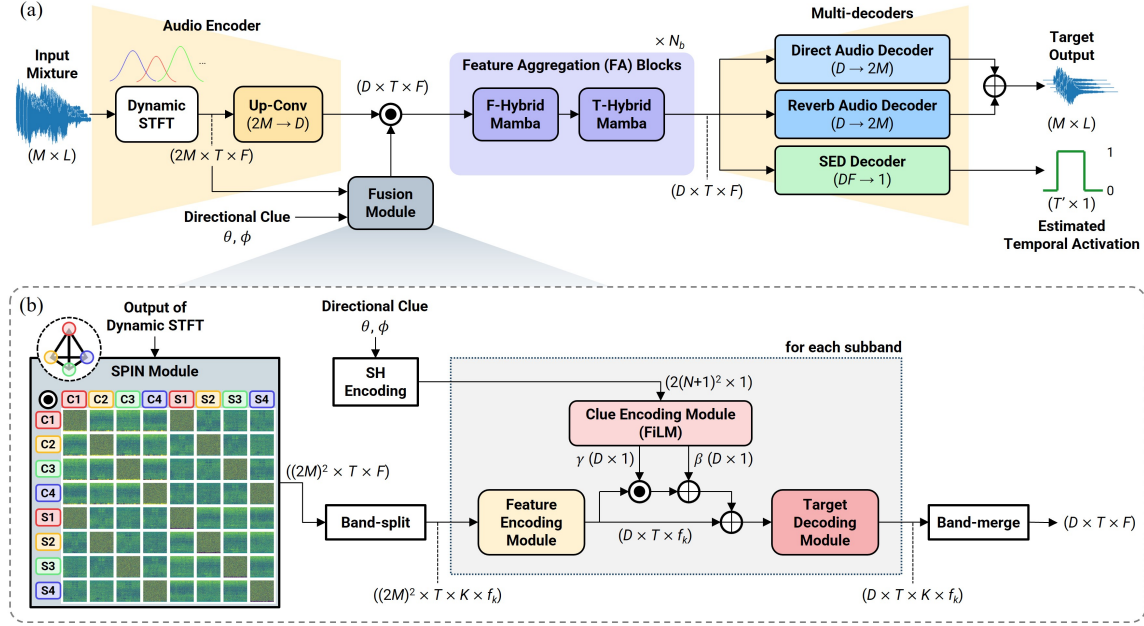
SPIN input feature To build a general input feature capturing full spatial information, we propose a Spectral Pairwise INteraction (SPIN) module that captures all pairwise interactions between sinusoidal components on complex spectrogram of multichannel signals. SPIN features are then fused with a DoA clue in overlapping frequency subbands to capture frequency-dependent spatial cues, extending the advantages of prior band-split approaches [19–22].

SH embedding for DoA clues We employ spherical harmonics (SH) as the DoA clue embedding, which provides a continuous angular representation across the 2D sphere. The SH embedding enables the model to handle any DoA value without discretization.

Iterative refinement with temporal clue Furthermore, we adopt an iterative refinement strategy inspired by the chain-of-inference (CoI) paradigm [23–25], where the estimated temporal activation is recursively fused with the DoA clue, enabling the model to improve separation quality under challenging multi-source conditions progressively.

Fig. 1. (a) Overall architecture of SoundCompass for DoA-based target sound extraction and (b) details of a fusion module including a Spectral Pairwise INteraction (SPIN) module and integrating directional clue by feature-wise linear modulation (FiLM) for K subbands.

2. SOUNDCOMPASS FRAMEWORK

2.1. Model Architecture

Our proposed model is based on DeepASA [25] backbone, which achieved state-of-the-art performance in universal source separation (USS) and sound event localization and detection (SELD). The architecture sequentially applies multi-head self-attention and Mamba feedforward networks along spectral and temporal dimensions separately to capture object-level features from mixtures. The original backbone separates object features without any clue, and our focus here is how to effectively guide the object separation process using a DoA clue.

The overall architecture for achieving this objective is illustrated in Fig. 1(a), excluding the batch dimension. A multichannel mixture is first transformed into a complex spectrogram of shape $2M \times T \times F$ using the short-time Fourier transform (STFT), where M, T, and F denote the number of microphones, time frames, and frequency bins, respectively. Instead of a fixed window, a learnable Gaussian window parameterized by adaptive mean and standard deviation is applied, allowing each frame to be spectralized from different temporal focus and spread. The spectrogram is then mapped into a feature space, increasing the channel dimension from $2M$ to D through a Conv2D encoder (kernel size 3 and stride 1) that extracts spatial cues and local spectral–temporal patterns. The resulting feature is modulated by a fusion module to align extracted features with the directional clue, as described in section 2.2. Subsequently, fused features are processed by feature aggregation (FA) blocks that analyze spectral and temporal dependencies and separate features corresponding to the directional clue. Finally, two audio decoders reconstruct the multichannel direct sound and reverberation separately by reducing the channel dimension from D back to $2M$ using Conv2D layers of kernel size 3 and stride 1, followed by an inverse STFT (iSTFT) to recover the target waveform.

2.2. Directional Clue Integration

Spectral Pairwise INteraction The fusion module takes the complex spectrogram and directional clues to generate the spatial feature mask that guides the DeepASA system to the target direction. To aid in extracting spatial details from the complex spectrogram, we present a SPIN module. Within this module, the cosine and sine components of individual microphone signals are multiplied in the complex spectrogram domain, yielding a channel dimension of $(2M)^2$. This multiplication enhances the recognition of inter-channel phase or time differences, as well as level differences when necessary. The sinusoidal products, confined within a range of ± 1, ensure stable learning dynamics during training. Given that inter-channel relationships often vary by frequency, we adopt a band-split strategy [19–22] to ensure that inter-channel features are developed and merged with directional clues in each frequency band. Specifically, we employ overlapping subbands based on the 12-TET Western musical scale [20, 22], using $K = 31$ subbands, with narrower bandwidths at lower frequencies and wider bandwidths at higher frequencies. This approach promotes continuity across overlapping frequencies and minimizes information loss at subband boundaries.

Spherical harmonics embedding As an accurate and continuous representation of the DoA clue (θ, ϕ), we employ spherical harmonics (SH) as embeddings, as depicted in Fig. 1(b). Unlike one-hot embeddings, SH embeddings allow for the representation of angles without the need for discretization. Additionally, in contrast to cycpos embeddings, which define azimuth and elevation angles separately, SH embeddings can describe the position on an S^2 sphere without coordinate separation, thus providing a consistent representation regardless of coordinate rotation. The complex spherical harmonics of order n and degree m are defined as

$$Y_n^m(\theta, \phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)e^{im\phi}, \quad (1)$$

where $P_n^m(\cdot)$ denotes associated Legendre functions. We use the

5-th order encoding by stacking the real and imaginary components of spherical harmonics, yielding an embedding vector of dimension $2(N + 1)^2$ for $N = 5$.

Fusion in subbands The encoded SH vector is fused with the output from the SPIN module in each subband. The clue encoding module is a FiLM [26] layer generating scale (γ) and shift (β) parameters for feature modulation, combined with a residual connection from the feature encoding module. This design enables fine-grained spatial conditioning without hand-crafted feature engineering. Each encoding and decoding block consists of a linear layer, adaptive layer normalization (AdaNorm) [27], and a parametric rectified linear unit (PReLU) activation.

Iterative refinement To enhance robustness, we incorporate a sound event detection (SED) decoder following the on/off decoder structure of DeepASA [25]. This module outputs a frame-wise binary sequence that indicates the presence of the target source at each time step. As illustrated in Fig. 2, this sequence is combined with the SH embedding to form a time-varying directional clue of shape $T' \times 2(N + 1)^2$, where T' denotes the sequence length of the SED decoder output. The clue is linearly interpolated to T in the time dimension and recursively injected into subsequent TSE stages. This chain-of-inference (CoI) strategy enables the model to iteratively refine its extraction by aligning directional information with temporal dynamics, with no additional modules required. During training, the already trained first-stage model is kept fixed, and only the subsequent stage is fine-tuned from the first-stage model using oracle time-varying clues while fixing its encoder and fusion module. This fine-tuned stage then functions as the subsequent stage. At evaluation, the complete pipeline is used: the output of the first-stage model is combined with the SH embedding and fed into the subsequent stage, which relies on SED predictions from the previous stage to progressively refine the separation of the target source.

2.3. Loss Functions

A linear combination of signal-to-noise ratio (SNR) and scale-invariant signal-to-noise ratio (SI-SNR) [28] loss with a ratio of 9:1 was used for the direct/reverb audio decoder and the sum of outputs from the decoders. Additionally, binary cross-entropy (BCE) loss was employed for the SED decoder to estimate the temporal activation of the target source. All terms were summed with identical weights.

3. EXPERIMENT AND ANALYSIS

3.1. Datasets

The proposed architecture was trained and evaluated on the Auditory Scene Analysis V2 (ASA2) dataset[1] [25], which contains 13 audio classes with 2–5 foreground sources and one background noise per mixture. However, to align with our direction-based TSE model, we regenerated the dataset with stationary sources using the gpuRIR library[2] [29], fixing each source at its initial position. Each mixture is 4 seconds long, sampled at 16 kHz, and recorded with a 4-channel tetrahedral microphone array of 4.2 cm radius centered in a cuboid room. The other configurations, including initial source positions and room reflections, follow those of the ASA2 dataset configuration. The resultant training, validation, and test sets comprise 50k, 2k, and 2k mixtures, respectively.
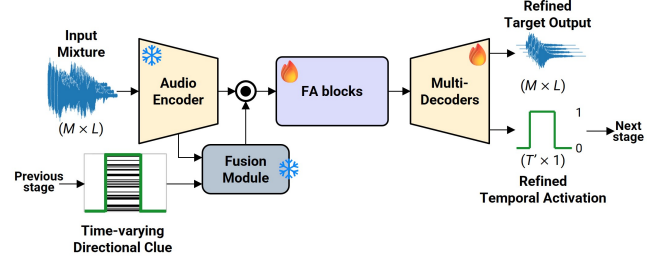

Fig. 2. Details of iterative refinement.

3.2. Training Setups and Evaluation Metrics

All training configurations largely followed [25], with several modifications. Optimization was performed using AdamW with an initial learning rate of 0.0005, which was reduced by a factor of 0.1 if the validation loss did not decrease for five consecutive epochs. Gradient norm clipping was applied with a threshold of 5, and training was conducted for 100 epochs with a batch size of 2 on four GeForce RTX 4090 GPUs.

The extraction performance was evaluated using SNR and SI-SNR improvements over the mixture. To further examine the spatial fidelity of multichannel extraction, the consistency of interchannel cues was assessed by computing the mean absolute error (MAE) between estimation and ground truth, including ΔILD, ΔIPD, and ΔITD across all microphone pairs. The ITD is derived using the generalized cross-correlation phase transform (GCC-PHAT)[3] [5]. All the above metrics were computed for each source in each mixture and then averaged. In addition, the model complexity was evaluated in terms of the number of trainable parameters (Param.) and total multiplications and additions (Mult-Adds)[4].

3.3. Analysis of Results

Table 1 presents comparisons to other TSE systems, as well as ablation studies of the proposed SoundCompass framework. First, the vanilla DeepASA model [25], representing universal source separation (USS) without any injected clue, achieves an SNRi of 15.6 dB and SI-SNRi of 13.0 dB. This unguided separation provides a baseline for evaluating the benefit of direction-aware extraction. To evaluate the benefit of DoA clue injection, we compared early and late fusion strategies. Injecting DoA clues before feature aggregation (FA) blocks using the fusion module consistently improves performance over late integration (after FA). This confirms the importance of exploiting spatial cues at early stages to achieve better target sound extraction and spatial fidelity.

We then compared SoundCompass with recent DoA-based single-channel TSE baselines, SSDQ[5] [12] and DSENet[6] [18]. For fair evaluation, we adapted their designs to match our setup of using one DoA clue: a point spatial query directing a point (θ, ϕ) instead of a region query for SSDQ and extended cyc-pos embedding to elevation as well as azimuth without beamwidth control for DSENet, following their training configurations. SSDQ, which relies on hand-crafted features, performs poorly in multi-source mixtures, while DSENet shows noticeable improvements but still falls short of SoundCompass. Notably, our method achieves higher SNR metrics
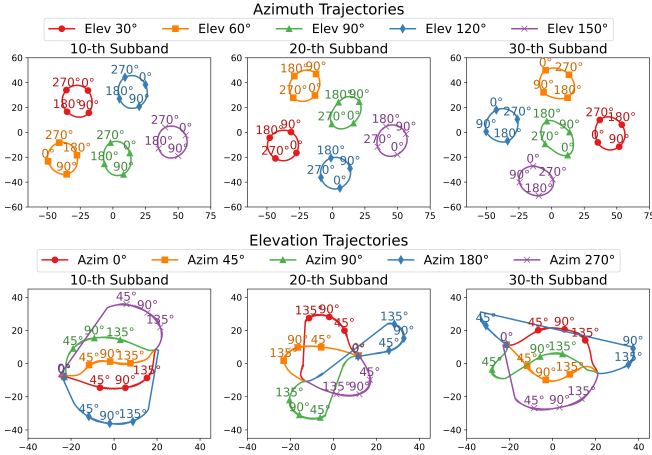

Table 1. Performance comparisons across models and structural variations of the proposed methods.

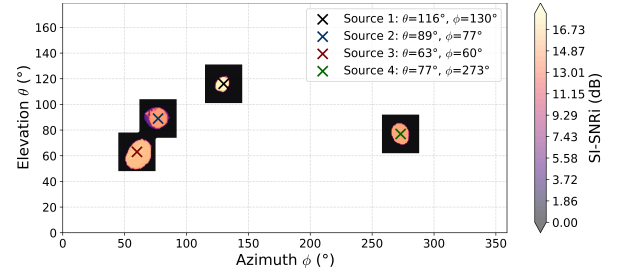| Model	SNR Metrics ↑		Spatial Errors ↓			Complexities ↓	
	SNRi (dB)	SI-SNRi (dB)	ΔILD (dB)	ΔIPD (rad)	ΔITD (μs)	Param.	Mult-Adds
Universal source separation							
DeepASA [25]	15.636	12.976	0.261	0.896	44.829	5.46 M	74.85 G
Target sound extraction							
SSDQ (w. point spatial query) [12]	5.949	-1.171	-	-	-	3.91 M	21.22 G
DSENet (w. cyc-pos (θ, ϕ)) [18]	16.419	16.025	-	-	-	4.88 M	86.89 G
Proposed (DoA after FA)	15.977	14.508	0.146	0.825	25.443	2.70 M	20.49 G
Proposed (DoA before FA)	17.865	16.717	0.099	0.805	10.302	2.70 M	20.49 G
remove an interaction in SPIN	5.663	15.854	0.115	0.821	11.765	2.59 M	20.49 G
replace SH to cyc-pos (θ, ϕ)	17.696	16.538	0.100	**0.782**	12.747	2.70 M	20.49 G
remove a band-split structure	17.524	16.238	0.104	0.808	14.513	2.16 M	20.49 G
add an SED decoder	17.884	16.780	0.098	0.800	9.993	4.09 M	23.46 G
refine iteratively ($\times$2)	**18.196**	**17.079**	**0.093**	0.789	**9.714**	+3.48 M	+24.01 G

Fig. 3. The t-SNE trajectories of the FiLM scale (γ) parameters across three subbands, with respect to azimuth (top, for 5 fixed elevations) and elevation (bottom, for 5 fixed azimuths).

Fig. 4. An example of SI-SNRi contour maps within $\pm 15°$ from each target direction marked as "X" in a cuboid room of size [width, length, height] = [5.57, 5.20, 3.79] m with an RT60 of 0.32 s.

while maintaining lower computational complexity, indicating both effectiveness and efficiency.

Ablation studies further highlight the contribution of each component. Removing pairwise interactions in the SPIN module (i.e., using only the raw $2M$ cosine and sine components without multiplication) causes a substantial degradation, highlighting the necessity of cross-channel correlation modeling. Replacing spherical harmonics (SH) with cyc-pos embeddings slightly degrades performance, while eliminating the band-split structure also reduces accuracy, underscoring the importance of frequency-dependent spatial cues. Incorporating an SED decoder shows modest improvements, while iterative refinement further boosts performance, demonstrating the advantage of progressively refining activations at the cost of additional parameters.

To better understand how the directional clue is embedded, Fig. 3 visualizes the t-SNE trajectories of the FiLM scale (γ) parameters from the clue encoding module across three subbands. For 5 fixed elevations, we projected the scale parameters across azimuths into the same feature space using t-SNE; the same procedure was applied to visualize variations across elevations for 5 fixed azimuths. Azimuthal variations form near-circular manifolds that remain distinct across elevations, indicating that angular periodicity is preserved. In contrast, elevation trajectories evolve from $0°$ to

$180°$ and converge to similar points across azimuths, reflecting the continuous nature of vertical cues. In addition, the different geometric patterns across subbands suggest that frequency-specific spatial correlations are captured, highlighting the benefit of band-split modulation.

Fig. 4 illustrates SI-SNRi sensitivity to deviations in directional clues. The contour maps show that performance peaks near the true source directions and degrades as the DoA estimate deviates by up to $\pm 15°$. Circular regions of high SI-SNRi form around each target position, demonstrating that SoundCompass framework effectively leverages directional guidance. Smaller regions indicate strong direction sensitivity, while broader regions suggest tolerance to some angular mismatch. This trade-off highlights the practical robustness of the proposed framework, as small DoA deviations are inevitable in real-world scenarios. The audio demo is available at https://choishio.github.io/demo-SoundCompass/.

4. CONCLUSION

We proposed SoundCompass, a DoA-based TSE framework that integrates spherical harmonics encoding with spectral pairwise interaction for efficient spatial conditioning. Through overlapping band-split modulation and sound event activation estimation, the model effectively captures both frequency-dependent and time-varying spatial cues with low complexity. Furthermore, the iterative refinement strategy highlights the advantage of coupling DoA clues with temporal dynamics, enabling a more robust extraction in diverse conditions. These results suggest promising directions toward more flexible source manipulation, such as handling dynamically moving sources by jointly estimating their time-varying DoA and activity.

5. REFERENCES

[1] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, Shanghai, China, 2020, ISCA, pp. 1441–1445.

[2] D. J. A. Padilla, N. L. Westhausen, S. Vivekananthan, and B. T. Meyer, "Location-aware target speaker extraction for hearing aids," in *Proc. Interspeech*, Rotterdam, Netherland, 2025, IEEE, pp. 2975–2979.

[3] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (spear) challenge," in *Proc. Int. Workshop Acoust. Signal Enhanc. (IWAENC)*, Bamberg, Germany, 2022, IEEE, pp. 1–5.

[4] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Process. Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[5] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic Hearing: Programming acoustic scenes with binaural hearables," in *Proc. 36th Annual ACM Symp. User Interface Software Tech. (UIST)*, San Francisco, CA, USA, 2023, ACM, pp. 1–15.

[6] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, and S. Araki, "Interaural time difference loss for binaural target sound extraction," in *Proc. Int. Workshop Acoust. Signal Enhanc. (IWAENC)*, Aalborg, Denmark, 2024, IEEE, pp. 210–214.

[7] H. Ma, Z. Peng, M. Shao, J. Liu, X. Li, and X. Wu, "CLAPSep: Leveraging contrastive pre-trained models for multi-modal query-conditioned target sound extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 4945–4960, 2024.

[8] H. Ma, Z. Peng, X. Li, Y. Li, M. Shao, Q. Kong, and J. Liu, "Language-queried target sound extraction without parallel training data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hyderabad, India, 2025, IEEE, pp. 1–5.

[9] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, K. Kashino, et al., "ConceptBeam: Concept driven target speech extraction," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 4252–4260.

[10] C. Li, Y. Qian, Z. Chen, D. Wang, T. Yoshioka, M. Zeng, et al., "Target sound extraction with variable cross-modality clues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes, Greece, 2023, IEEE, pp. 1–5.

[11] R. Gu and Y. Luo, "ReZero: Region-customizable sound extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 2576–2589, 2024.

[12] X. Zhu, X. Qian, and D. Liang, "SSDQ: Target speaker extraction via semantic and spatial dual querying," *IEEE Signal Process. Letters (SPL)*, vol. 32, pp. 3167–3171, 2025.

[13] S. Zhang, J. Zhang, Y. Wang, and H. Yan, "DOA or Speaker Embedding: Which is better for multi-microphone target speaker extraction," *IEEE Signal Process. Letters (SPL)*, vol. 32, pp. 3350–3354, 2025.

[14] A. Pandey, S. Lee, J. Azcarreta, D. Wong, and B. Xu, "All neural low-latency directional speech extraction," in *Proc. Interspeech*, Kos, Greece, 2024, ISCA, pp. 4328–4332.

[15] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 542–553, 2023.

[16] Y. He, A. Markham, and O. Köpüklü, "SoundTRC: Dnn-based acoustic target region control," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hyderabad, India, 2025, IEEE, pp. 1–5.

[17] D. Choi and J. W. Choi, "Multichannel-to-multichannel target sound extraction using direction and timestamp clues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hyderabad, India, 2025, IEEE, pp. 1–5.

[18] K. Jing, W. Zhang, and Y. Gao, "End-to-end doa-guided speech extraction in noisy multi-talker scenarios," in *Proc. Interspeech*, Rotterdam, Netherlands, 2025, ISCA, pp. 1443–1447.

[19] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 31, pp. 1893–1901, 2023.

[20] K. N. Watcharasupat, C. W. Wu, Y. Ding, I. Orife, A. J. Hipple, W. Wolcott, et al., "A generalized bandsplit neural network for cinematic audio source separation," *IEEE Open Journal of Sig. Process. (OJSP)*, vol. 5, pp. 73–81, 2023.

[21] Y. Yang, H. Li, X. Wang, W. Zhang, S. Makino, and J. Chen, "Stereophonic music source separation with spatially-informed bridging band-split network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seoul, Korea, 2024, IEEE, pp. 786–790.

[22] K. N. Watcharasupat and A. Lerch, "A stem-agnostic single-decoder system for music source separation beyond four stems," in *Proc. 25th Int. Soc. for Music Inform. Retriev. Conf. (ISMIR)*, San Francisco, CA, USA, 2024.

[23] D. Wu, X. Wu, and T. Qu, "Leveraging moving sound source trajectories for universal sound separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 33, pp. 2337–2348, 2024.

[24] Y. Kwon, D. Lee, D. Kim, and J. W. Choi, "Self-guided target sound extraction and classification through universal sound separation model and multiple clues," Tech. Rep., DCASE 2025 Challenge, June 2025.

[25] D. Lee, Y. Kwon, and J. W. Choi, "DeepASA: An object-oriented one-for-all network for auditory scene analysis," in *Proc. Advances in Neural Inform. Process. Systems (NeurIPS)*, San Diego, USA, 2025, Curran Associates, Inc.

[26] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, 2018, AAAI Press, vol. 32, pp. 3942–3951.

[27] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," in *Proc. Advances in Neural Inform. Process. Systems (NeurIPS)*, Vancouver, Canada, 2019, vol. 32, pp. 4383–4393, Curran Associates, Inc.

[28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, 2019, IEEE, pp. 626–630.

[29] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Appl.*, vol. 80, no. 4, pp. 5653–5671, 2021.